Response to Referee 2

Review of

Retrieval of thermodynamic profiles in the lower troposphere from GNSS radio occultation using deep learning

by Aichinger-Rosenberger and Sjoberg

General comments

The manuscript presents an artificial neural network (ANN) approach for the retrieval of temperature, pressure, and specific humidity profiles in the neutral atmosphere, using a proposed framework called Advancing the GNSS-RO retrieval of atmospheric profiles using MAchine-learning (AROMA). Model training is based on a large dataset of profiles from COSMIC2, commercial data from Spire, and CDAAC 1D-Var products, which serve as target values for pressure, temperature, and specific humidity. Validation is carried out against 1D-Var profiles from CDAAC (set not used during training), as well as ERA5 reanalyses, radiosondes, and commercial data from PlanetiQ receivers. While the authors report generally small errors and high correlation values when comparing model outputs to validation datasets, the analysis of the results lacks depth and the comparison to previous studies remains unclear.

I believe a comparison against previous studies is crucial. Including RMSE results would facilitate comparison with works such as Lasota et al. (2021) (while acknowledging differences between the studies). Although some quantitative values are provided, I find the analysis of results somewhat vague. The manuscript frequently uses terms like “agree very well,” “very high agreement,” and “similar overall performance” without specifying the range of agreement or indicating whether the differences are statistically significant. An assessment of which specific regions may be performing better or worse could also be beneficial.

In addition, the authors state that a key advantage of their approach is its independence from external meteorological data. The model is trained using 1D-Var products as reference data, which are themselves based on ECMWF background fields, meaning that the training data is not completely independent. The authors comment on this in the Conclusions section but still claim the AROMA’s main advantage is the independence of external meteorological data. I think this limitation should be more clearly acknowledged and articulated in the manuscript.

Overall, the paper is well written and well structured. I think it is encouraging to see incremental progress in the application of machine learning techniques to GNSS-RO data. The study is relevant and contributes to the field by expanding the training dataset in terms of size and time coverage, incorporating commercial GNSS-RO data for training and validation. However, I think the presentation and discussion of the results needs to be improved and the novelty better articulated. Therefore, I suggest major improvements are needed before publication.

Thank you very much for your review and feedback, as well as all the valuable comments and suggestions on how to improve the manuscript. We will try to extend and strengthen the performance analysis carried out, based on these suggestions.

One important aspect of our approach is its level of independence from external meteorological data, especially in comparison to the classic 1D-VAR technique. Here we want to clarify the misconceptions which have shown up in the reviews.

We never claimed that our approach is fully independent of external data, anywhere in the manuscript. This is obvious, since the model is trained using the operational 1D-VAR results. However, its level of independence is significantly higher than for 1D-VAR, as it does not need any auxiliary data to run the model once it is trained. The level of independence is determined by the amount of data used during model training. This tradeoff also explains why limiting the amount of training data is a reasonable approach, although much larger data sets could be utilized. We will try to clarify this further in the revised manuscript.

Regarding quantitative language used in the manuscript, we agree that the frequent use of subjective descriptors is not appropriate. We will amend relevant statements that we find during our revision.

In the following, we have answered specific comments one-by-one.

Specific comments

L10: “from both internal and external validation”. Is this a common terminology for ML studies?

We will change the wording throughout the whole manuscript to e.g. *Performance evaluation on (test data / ERA5 /RAOB)*

L24-26: The study explores an alternative approach to retrieving thermodynamic profiles from RO; however, the importance and practical utility of these profiles are barely discussed.

We will add information on various use cases of thermodynamic profiles

L54: The term “internal and external validation” is introduced for the first time in the main body of the manuscript. Apologies if I am unfamiliar, but this doesn’t seem to be a commonly used terminology. If authors choose to keep it, it would be helpful to briefly define it here for clarity.

We will change the wording throughout the whole manuscript to e.g. *Performance evaluation on (test data / ERA5 /RAOB),* as also suggested by other reviewers.

L62: How similar or different is your ANN compared to previous studies?

The ANN used in this study is quite similar to what was used by Lasota (2021) and Hooda et. al (2023). However, we use a different set of hyperparameters and features on which the model is trained. We will comment on this in the revised manuscript.

L67: I recommend including a brief outline of the paper at the end of the Introduction section. This would help guide the reader and clarify the flow of the manuscript, especially given the multi-step nature of the proposed methodology.

This outline will be added to the manuscript.

L115: RAOB is defined twice in the manuscript but is not used consistently throughout the document. Please revise.

This will be corrected in the revised manuscript.

L118: Could the authors clarify the rationale for choosing a 500 km collocation distance for RO observations in this study? This value appears notably larger than the 200 km distance that is typically used in the literature.

The main motivation was to have a larger sample size for the chosen period. We know this threshold is large and might introduce representativeness errors. However, these errors would identically impact both the 1D-Var and AROMA error statistics. The actual absolute errors are not the most important information communicated in these comparisons. The focus of our interpretation is more on the relative differences between 1D-Var and AROMA, which are small to non-existent. Identical error levels as 1D-Var is, by definition, the best possible result to be achieved with our target setup.

Still, the representativeness errors may be overwhelming the actual differences between 1D-Var and AROMA, so we plan to extend the amount of RAOB observations included in the validation by extending the validation period. Then we will be able to have an equal (or larger) sample size while reducing the collocation distance to the tyipcal 200-300 km.

L136: Please provide a citation for the constant terms.

We will add a citation in the revised manuscript.

L164: With regard to the ANN, can you explain why you chose a feed-forward multilayer perceptron (MLP) over more modern alternatives? Could more complex or structured architectures (e.g., CNNs, RNNs) be better suited?

The focus of this study was not to necessarily explore the performance of new, more complex deep learning architectures, but rather to investigate the impact of additional data, as provided by commercial RO missions. Therefore, we stuck to the simple ANN architecture which already showed some promising results in previous studies. Nevertheless, we are currently exploring more complex models for the same task, but these investigations are out of scope of this work.

L165: The statement “ANNs are supervised neural networks,” is not correct. Please revise to reflect that ANNs can be used in both supervised and unsupervised learning contexts.

This statement will be revised.

L175: With regard to hyperparameter tuning, this seems to be slightly more explored in Section 3.3 on model setup. However, I think more context should be provided on how this is done and what is the practice in other studies.

L182-188: In general, there is a lack of justification for the chosen ANN architecture and the hyperparameter tuning process. What is the reasoning behind the parameter values presented in Table 1? Would it have been feasible to test larger batch sizes or a greater number of epochs? What limitations have you encountered? It would be helpful if the authors could comment on why this particular combination was the most successful, as this insight could be valuable for future work in this area. Additionally, providing figures or metrics to support these results would strengthen the manuscript.

Regarding the last two points: We will extend the number of combinations we test for hyperparameter tuning and the documentation of RMSE and correlation metrics for those setups. Therefore, a small grid search will be carried out. Extensive hyperparameter tuning is not doable in this study because of computational and time resources.

L194: signal-to-noise ratio (SNR) is defined in L74. Please use it accordingly.

We will correct this in the revised manuscript.

L227: Is there a reason why the top height level is 20 km? Climate and other studies using the retrieved thermodynamic RO profiles use data only up to this height?

The reason is that for this work, the focus is on the lower atmosphere. We plan to extend this up to 60 km altitude in future studies, as for the profiles typically provided by CDAAC in the wetPf2 format.

Section 4: As noted in my general comments, it would be helpful to use a more specific metrics range instead of vague terms. Providing actual ranges for the reported agreement would make the analysis clearer and more informative.

We will reformulate the relevant sections and provide the actual ranges for the agreement between AROMA and the respective validation data source.

L259: There is a negative bias above 12 km in both temperature and pressure in AROMA. Is this observed in other studies as well? Can the authors comment on what could be causing these biases? Are there specific regions contributing to them?

Thank you, this is an interesting point. We did not investigate these biases in detail during this study, but we'll try to do so in the revised manuscript.

Figure 1, 2, 3, 4, 6, and 7: All these figures could benefit from adding a letter identification. Also, I think these plots would benefit from adding the RMSE profiles in addition to or instead of the STD. Confidence intervals would also be helpful to see in these plots. What binning size is used?

Thank you for these comments.

We plan to add these letter identifications to the revised manuscript.

We also plan to add RMSE to figures and tables.

On significance testing, we don't agree that demonstrating whether the differences between the CDAAC and AROMA methods of moist atmospheric retrieval are significantly different (or not) adds much useful information to the reader. There are 1.5 million input profiles to these STDV values. Even by subsampling only 1% of all profiles in each dataset, the criterion for a 5% significant difference is a relative difference of ~1% between the two STDV values at any level. Given Figures 2-4, we are therefore likely to find that the STDV differences are significantly different everywhere. This does not mean that AROMA does poorly, only that we are confident that the standard deviations are different.

Vertical step size in these figures and analyses is 100 m.

Table 4: not referenced in the manuscript. Please revise.

Thanks for the hint, Tables 4 and 5 will be referenced.

L331: "instabilities in the retrieval." Can you clarify what this means?

Although we have not investigated this in detail, we suspect that it is observation noise learned by the model, in an altitude region where on average only small amounts of moisture are present. We will elaborate in the revised manuscript.